# You Don't Know Data
# (or Information)

Willam C. Burkett

December 2023

W. Edwards Deming, the father of statistical quality control, said:

*"If you can't describe what you are doing as a process, you don't know what you're doing".*

When looking at the world of IT and applied to the dichotomy of *software* and *data*, Deming's quote applies the software part of that pair. The analogous principle[1] that applies to the data side of the pair is:

*If you can't define it, you (probably) don't really know what it is.*

I've been a data modeller for four decades and over the course of my career have consistently pursued answers to the question: what makes a data model good? That pursuit has inexorably led me to the critical importance of good definitions. Sound, clear, unambiguous definitions are not only a valuable part of data model documentation, but they also force a data modeller like me to ***really*** understand and be clear about the domain I am modelling and how I am modelling it – it forces me to "really know what it is.". Simply: the process of crafting good definitions makes the data model better.

The purpose of this essay is to apply a little of what I've learned about defining things to two terms that are very fundamental to our practice of data management and frequently appear conjoined in data management literature: "data and information". Personally – and maybe it's just me – I find the use of "data and information" as a pair to be a cop-out. By using "data and information", is the author just trying to cast a wide net of meaning and "cover all the bases" without doing the work to ascertain whether they mean "data" or "information"? The fact that they include both implies that they do see some distinction between them – but what is that distinction?

In order to answer that question – and in the process, perhaps, shed some light of the vague use of language that plagues our field of work – we must first consider the question: What makes a good definition?

## Crafting High-Quality Definitions

In his lexicon of software requirements and specifications[2], Michael Jackson (the British software consultant, not the other guy) described *designations* as a tool for unambiguous (or less ambiguous) identification and naming of things. A *designation* is comprised of a

[1] A generalized statement of position that is accepted as true or valid, and often reflects values, beliefs, or convictions on the "right" or "best" way to do or achieve a result.

[2] Jackson, M. (1995). Software Requirements & Specifications: A Lexicon of Practice, Principles and Prejudices. New York, ACM Press.

*designated term* (or definiens, if you'd like to use lexicographic terminology) and a set of *recognition rules* (definiendum). For example:

> Designated term: Duck(*x*)
>
> Recognition Rules:
>
> > *x* looks like a duck
> >
> > x walks like a duck
> >
> > *x* talks like a duck

The recognition rules are predicates that pick out properties of the thing we're trying to define and are used these to recognize and name the thing.

Similarly, *ISO 704 Terminology work — Principles and methods*[3] presents guidelines for defining concepts based on characteristics of those concepts:

> "A definition shall define the concept as a unit with a unique intension and extension[4]. The unique combination of characteristics creating the intention shall identify the concept and differentiate it from other concepts. The quality of most terminological products will be determined by the quality of the definitions."

*Terms* - which is what the standard is about - are human-created, language-based labels "attributed to the concept". A *term* is a word, phrase, name, or symbol we use to refer to the concept. ISO 704 also describes kinds of relationships between concepts and the development of "concept systems".

If that sounds a little be like data modelling to you, then I'd like to encourage and reinforce that thought. Even though ISO 704 says nothing about data modelling, I personally consider it the best "data modelling book" I've ever read. Data models are definitely "terminological products." But I'm biased, of course, because of my previously-stated position on the importance of high-quality definitions in the development of high quality data models.

The key element of both of these guidelines is the identification of unique and unambiguous properties (characteristics) of the thing being defined and use these properties both to define and differentiate *x* and *y*.

## Data and Information are Different Things

Before we explore what makes "data" and "information" different, let's first consider the *kinds* of words that they are. This is important because it affects both their definition and use.

"Data" and "information" are both nouns – that much is obvious. But have you ever heard anyone say things like "Those two data over there" or "Give me five informations"? Of course you haven't. Such phrases are not grammatically correct because *data* and *information* are **mass** nouns like the words *air*, *water*, and *sand*. You can say "two gigabytes of data" just like you say "two gallons of water" by attaching a measure to the mass noun, indicating an amount of it. I don't know of an analogous measure one can use with *information* – we speak of "information overload", meaning "lots of information", but "lots" is a very ambiguous kind of measure.

[3] ISO (2000). ISO 1087-1 Terminology Work - Vocabulary - Part 1: Theory and Application. Geneva, International Standards Organization (ISO).

[4] https://en.wikipedia.org/wiki/Extensional_and_intensional_definitions

**Count** nouns – as the name overtly states – are things that are countable: two baseballs, five hard drives, 10 books, 50 files, 5 golden ringssssss. Counting "data" requires attaching some kind of objective measure to it, like gigabytes, or picking out countable things like hard drives or files that contain some amount of data. There is no objective way to "count" information, as far as I know.

As mentioned in the introduction, the frequent use of "data and information" as a pair implies that the author does see something different about them, even if they can't put their finger on it and choose one or the other. Our ability to find ways to "count" data and inability to "count" information highlights one way in which they are different.

That they are, in fact, different can be easily demonstrated. Suppose that Pat knows that the range of an HG missile is 1000 nautical miles (Figure 1). Pat can express and convey that information in any number of ways:

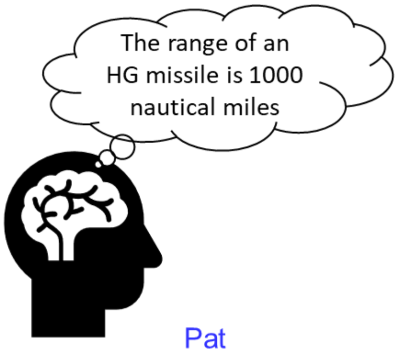


Figure 1

- By speaking
- By writing
- By drawing a picture
- By typing an email
- By structuring values in a database

Each of these is a mode or manner of expression (Figure 2). Regardless of the choice of mode/manner, Pat is expressing the *same information*. [5]

"Speech" is a mode of expression, as is writing and drawing. If Pat were to actually say "The range of the HG long-range missile is 100 nautical miles" to Chris at 4:55 pm on November 4th, 2023, that would be an instance of a "speech act" or an "utterance". We could call the other instances of using different modes of expression illustrated in Figure 2 "speech acts" or "utterances" but that would be a little awkward or misleading since we naturally associate the words "speech" and "utterance" with auditory use of natural language. So instead, let's call them "information artifacts" and define that term using the Jackson-style recognition rules:

The defining property of Artifact (x) is

*x* is produced by human

The defining properties of Information Artifact (y) are

*y* is an Artifact

*y* encodes information for the purpose of sharing or storage

(Notice the additive nature of these definitions.)

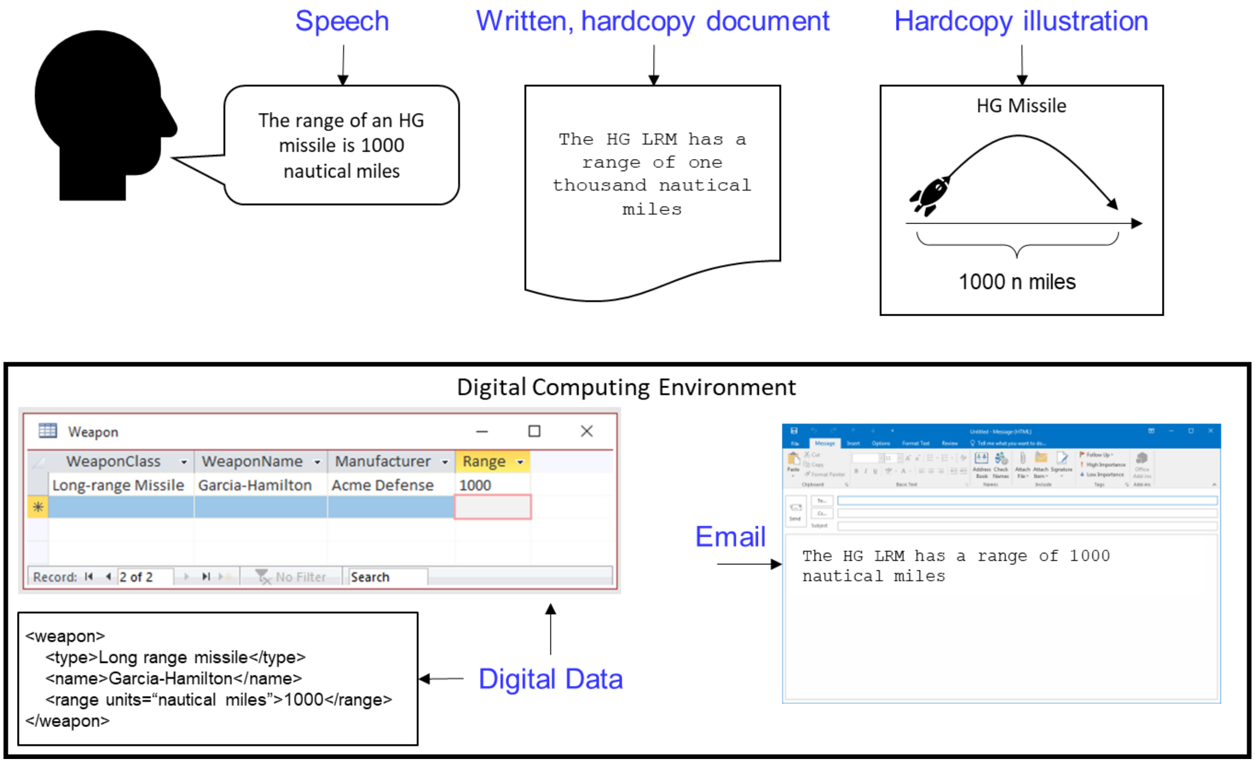
Admittedly, "information artifact" is a little awkward to use, too. It does, however, have some advantages to the present discussion and the overall understanding of the "data" and "information"; the term "information artifact" is:

- clearly about "information"

[5] There are minor informational differences between the expressions, but we shall set these aside for now.

- discrete and countable
- the intentional product of human action
- concrete, tangible, and objectively perceivable

In other words, an information artifact is an object that exists in the real world that is objectively perceivable by any human being (perhaps indirectly through the use of instruments) and contains information intentionally "put there" by a human being (or instrument created by a human being).


Speech
Written, hardcopy document
Hardcopy illustration
The range of an HG missile is 1000 nautical miles
The HG LRM has a range of one thousand nautical miles
HG Missile
1000 n miles
Digital Computing Environment
Weapon
Record: 2 of 2
No Filter
Search
Email
Digital Data

| WeaponClass | WeaponName | Manufacturer | Range |
|---|---|---|---|
| Long-range Missile | Garcia-Hamilton | Acme Defense | 1000 |

```
<weapon>
   <type>Long range missile</type>
   <name>Garcia-Hamilton</name>
   <range units="nautical miles">1000</range>
</weapon>
```

The HG LRM has a range of 1000 nautical miles

Figure 2

## Data

Now let's come back to the word "data". Using "data" as a mass noun is fine for generalized conversation, similar to "the air we breath" or "the water we drink". But in order to *manage* data and – critically – *objectively* manage data *as an asset* we have to find a way to use "data" as a count noun. We need a way to define the objective, countable, and unambiguous "data things" we are managing.

Asset (x) :=

*x* is objectively and unambiguously *trackable*

Information artifacts are discrete, countable, and objectively trackable; they are, therefore, "assets". Instead of "information artifact" we can synonymously call the things illustrated in Figure 2 "information assets".

Are information assets "data" or do they contain "data"? It depends on where you want to draw the line. The items found within a digital computing environment in Figure 2 are commonly called "data".[6] Where you draw your line depends on whether you want to consider the sound waves of speech or the marks on paper as "data".

Personally, I favor recognizing and acknowledging that we are all in the IT business and when we use the term "data" we are almost always, 99.99% of the time, referring to structured bit patterns in a computing system. Therefore, let's define "data" as follows:

Data (x) (IT):

> *x* is structured patterns of binary digits (bits, "1's and 0's") within digital computing technology
>
> *x* is or may be processed by software applications
>
> *x* is created in accordance with a specification to encode information which asserts the meaning of the structured bit patterns
>
> *x* is persistently stored on digital media, held in computing memory, or serialized in a transmitted message/data packet.

If, on the other hand, you are more inclined toward a more general definition of "data" that include sound waves, marks on paper, or analog patterns on magnetic tape:

Data (x) (Communication):

> *x* is a collection of physical symbols that are drawn from and created in accordance with a language with the intention to encode and represent information and manifested in a physical artifact.

The later definition encompasses the former insofar as the "structured bit patterns" are the physical symbols that exist on/in a physical artifact.

These definitions help us a little when it comes to data management, but, unfortunately, not quite enough. These definitions leave "data" as a mass noun – untrackable. To refer to data that is trackable, let's combine "data" (in the structured bit-pattern sense) and "asset":

Data Asset (x):

> *x* is a container, i.e., x has clear, unambiguous, and definitive boundaries; something is either "in" or "not in" the container
>
> *x* contains or consists of data (structured bit-patterns)
>
> *x* has a single unique, holistic identifier (e.g., serial number, path/file name, database ID, message ID)
>
> *x* is governed by a single data specification (e.g., a schema/physical data model), or set of integrated data specifications

Other things that are true of a data asset that may not necessarily be considered as part of the definition but are nonetheless important:

> All data type or data structure names for the data elements (or other compositional components of the data) within *x* are unique. This is enforced/entailed by the single data specification. A name "means" one thing both semantically and structurally.

[6] The relational table and XML element are often referred to as "structured data" and email (and documents) as "unstructured data".

> The boundaries of *x* (the container) define a managed identifier space within which all identifiers (e.g., relational data keys, record IDs, or XML "ID" attributes) are unique.

These definitions could be generalized to the term "information asset" (or "information artifact") if you prefer to start with the communication-context definition of "data"

## Information

At this point, we've pretty much nailed down the definition of "data". Now let's turn to "information".

> Data is a *means* …
>
> > … to represent (encode) information
> >
> > … to share information
>
> Communication (i.e., "sharing information") is the *end*

"Information" is not tangible and is inextricably linked to the creator[7] of the data (or information artifact) and, *separately and independently*, to the interpreter of the data. I use the words "separately and independently" to highlight the fact conveyed by the adage: "the message sent is not always the message received". On one hand, the skill with which the creator formulates and objectively expresses the "data" directly and materially impacts the effectiveness of conveying the intended information to an audience (interpreters). This is why those preparing a presentation are told to "know your audience." And this is where and why badly designed data models cause so many problems in information system implementations. Look back at the information represented in the relational database table in Figure 2 - there are two crucial bits of information missing from this particular information artifact that are overt in the others: (1) the fact that the range of the missile is stated in *nautical* miles; and (2) the units of measure for the value of "1000" in the column "Range". Both of these bits of information may (and should) be stipulated in the data model governing this table, but what if the data model isn't documented? (Which, as we *allllll* know, is far far far too often the case.) Perhaps the column should be named "range_in_nautical_miles".

On the other hand, different interpreters may obtain or extract different information from the same data. The information they obtain by "reading" the data depends on situation, context, motives, and, crucially, the background knowledge they bring and apply to the interpretive process. "A picture is worth a 1000 words."

So, one way to look at "information" is as what creators "put into" data and what interpreters "take out of" data. This is my preferred way to think about and use the word "information".

> Information (x):
>
> > *x* is that which is "put into" data (or "encoded as" data) by a creator for the purpose of conveying knowledge.
> >
> > *x* is that which is "taken out of" data (or "decoded from" data) by an interpreter for the purpose of learning (i.e., receiving knowledge.)

[7] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/elements11/creator/

Despite my best efforts, this definition is still “squishy” and I’m not sure how to objectively improve it. It is also important to note that this definition is different than but compatible with Claude Shannon’s definition of “information”[8].

It is entirely reasonable to discuss “the information contained in” a data asset or information artifact, but keep in mind the caveat that that “information” is never definitive, finite, or absolute. The only exceptions to this claim are very small data assets – for example, a bit in a digital computer is either “0” or “1”. A bit can only contain two “informations”[9]: (1) zero/false/no and (2) one/true/yes. (This is in keeping with Shannon’s definition of ”information”, which is really “information capacity”.)

The definition “information” connects the “data and information” question to an often-seen trio of terms: “data, information, and knowledge”. “Knowledge” is also an uncountable mass noun. My understanding and use of the term “knowledge” is the common understanding of the term in regular use: “what someone knows” – that is: “what is in, and what’s going on in, their brain”. Limiting “knowledge” to what is in someone’s brain obviously leads to a problem with terms like “knowledge-bases”; it is my opinion that there are no such things as “knowledge-bases”. “Knowledge-base” is a just a decorative and sales-y name for databases that purport to “be like a brain” insofar as the information they contain and processing capabilities enabled by the data structuring paradigm are “brain-function-like”. It is just an example of anthropomorphizing computing technology and I myself consider “knowledge-base” and even “artificial intelligence” just to be very clever data processing – because that’s exactly what it is.

The relationship among data, information and knowledge is illustrated in Figure 3. You can tell by the primitive graphics used in this illustration how long I’ve held and promulgated these understandings of data, information, and knowledge.

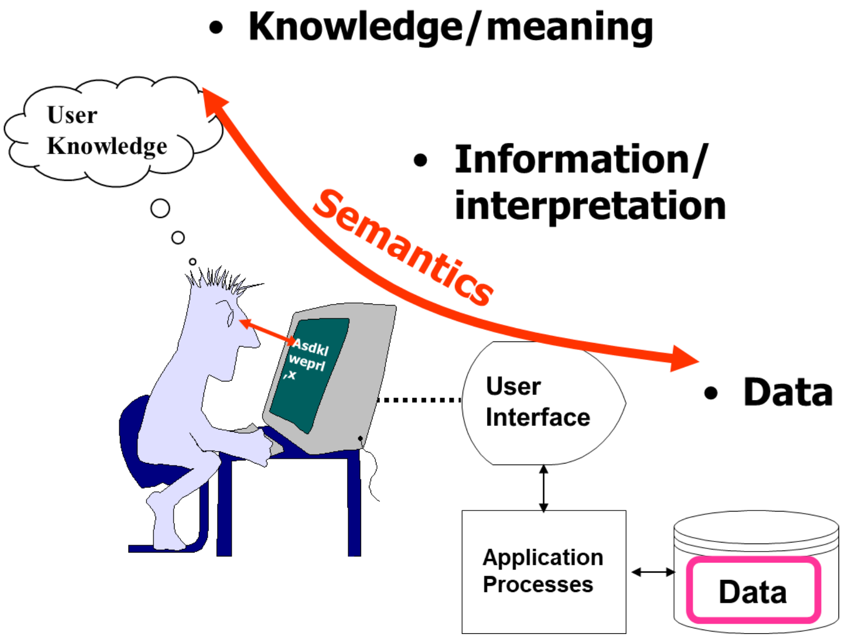


Figure 3

[8] https://en.wikipedia.org/wiki/Information_theory
[9] I use the term “infomeme” to refer to a single, distinct, non-decomposable piece of “information”. Multiple infomemes can be contained in a single data element.

## Ramifications: Data Management or Information Management?

The distinction between data and information has distinct impacts on the discipline of data management. First and foremost is the recognition that many data management disciplines conflate data management and information management. For example, Master Data Management for Customers: managing customer data and ensuring its quality and availability is a surrogate for managing customer information. Managing customer *information* is what is important; managing the data that conveys this information is the engineered, mechanized *means* that implements customer information management requirements.

On the other hand, backups, files, paths, packets, and checksums are clearly just *data* management. They are activities associated with data assets that don’t care about the information “in” the data assets. The data asset is, from this perspective, merely a commodity that is managed with logistical processes, like those used by UPS, USPS, and Fedex. But even then you can’t get away from *some* meaningful data: the data that is “slapped on” to these commoditized data assets – commonly called *metadata* – certainly contain information and are meaningful to the logisticians tracking, controlling, and moving the data assets around.

Is this distinction important? I’m not sure. I do think it’s important to understand and document the requirements for managing information (such as customer information):

- Who’s the authority for asserting this information is correct and that information is inaccurate?
- What individual pieces of information are important?

And it is important to do this before a single data structure is designed to represent/encode this information. It’s the same thing as determining what your business or endeavor needs to accomplish (functional requirements) before buying any tools to pursue those objectives.

I do think the distinction is valuable in differentiating the “ends” of data management capabilities. Master Data Management is really about managing information. Data Quality involves both *data* quality (syntactic correctness) and *information* quality (accurate meaning). Metadata management draws a sharp line between a data asset as a commodity (don’t care about the information “inside” the asset) and data *about* the commodity (meaningful to the commodity manager.) Data modelling has *lllloooonnngggg* been plagued by competition, tension, and conflict between two ends: specifying the physical structure of *data* (the role of a “physical data model”) and representation of the information required for enterprise processes and purposes (the role of a conventional “conceptual data model” – although the term “conceptual data model” is a harmful misnomer because it’s not really a model of “data”).

## Recap

The importance of clear, unambiguous definitions to the field of data management, and data modelling in particular, cannot be overstated. This is particularly evident in the way that we casually use the words “data” and “information” without ever quite exactly knowing what we’re referring to. Are we talking about structured bit patterns in a digital computing system, or are we talking about the information represented by those bit patterns? They are definitively not the same things, but our casual use of the terms often conflates their meanings and adds to our confusion about what we’re really talking about. The goal of this paper was to present a simple method for crafting clearer, less ambiguous definitions and apply that to the analysis of what kind of words “data” and “information” are and what they really mean.

Goal was met (x)[10]

*x* method(s) for crafting clearer definitions provided

*x* the word “data” analyzed and defined

*x* the word “information” analyzed and defined

*x* sound argument made that “data” and “information” are not the same thing

[10] This is not really a definition – just a fun way to conclude the article.